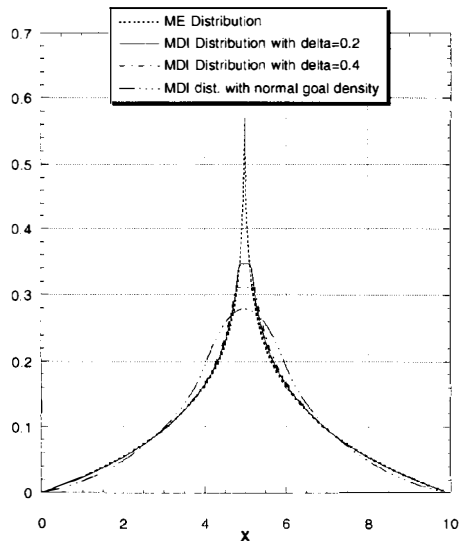

Fig. 7.

straints in several areas. Unfortunately, these estimates may not yield unimodal densities. This correspondence presents a method for transforming the estimation problem in the case of unimodal density estimation. The transformed problem is then solved by information-theoretic methods and transformed back to obtain a unimodal density estimate. Other qualitative characteristics of the desired density such as smoothness near the mode can also be incorporated into this unimodal information-theoretic density estimation technique.

REFERENCES

[1] J. P. Burg, "Maximum entropy spectral analysis," Ph.D. dissertation, Stanford Univ. Stanford. CA, 1975. (University of Microfilms No. AAD75-25,499)

[2] E. Parzen, "Quantile, parameter-select density estimation, and bi-information parameter estimators," in *Proc. NASA Workshop on Density Estimation and Function Smoothing* (Texas A&M Univ., College Station, TX, 1982), pp. 60–84.

[3] J. H. Kemperman, "Moment problems with convexity conditions I," in *Optimizing Methods in Statistics*. New York: Academic Press, 1971.

[4] P. L. Brockett, A. Charnes, L. Golden, and K. H. Paick, "A Method for obtaining a unimodal prior distribution density," CCS Report 473, Center for Cybernetic Studies, The Univ. of Texas at Austin, 1983.

[5] N. Wiener, *Cybernetics*. New York: Wiley, 1948.

[6] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 76–86, 1951.

[7] H. Akaike, "An extension of the model of maximum likelihood and Stein's problem," *Ann. Inst. Statist. Math.*, vol. 29, pt. A, pp. 153–164, 1977.

[8] ——, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. on Information Theory*, B. N. Petrov and F. Csaki, Eds. (Budapest, Hungary, Akademiai Kiado, 1973), pp. 267–281.

[9] D. V. Gokhale and S. Kullback, *The Information in Contingency Tables*. New York: Marcel Dekker, 1978.

[10] S. Guiasu, *Information Theory with Application*. London, UK: McGraw-Hill, 1977.

[11] P. L. Brockett, A. Charnes, and W. W. Cooper, "MDI estimation via unconstrained convex programming," *Commun. Stat. Simul. and Comput.*, vol. B9, no. 3, p. 223, 1980.

[12] A. Charnes, W. W. Cooper, and L. Seiford, "Extremal principles and optimization qualities for Khinchin–Kullback–Leibler estimation," *Math. Operationsforsch. Statist. Ser. Optimization*, vol. 9, no. 1, pp. 21–29, 1978.

[13] W. Feller, *An Introduction to Probability Theory and Its Application*, vol. 2. New York: Wiley, 1971.

[14] T. Sager, "Consistency in non-parameteric estimation of the mode," *Ann. Stat.*, vol. 3, pp. 698–706, 1975.

[15] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[16] L. S. Lasdon, A. D. Waren, A. Jain, and M. Ratner, "Design and testing of a generalized reduced gradient code for non-linear programming," *ACM Trans. Math. Software*, vol. 4, no. 1.

On-Line Density Estimators with High Efficiency

Ola Hössjer and Ulla Holst

***Abstract*—We present on-line procedures for estimating density functions and their derivatives. At each step, M terms are updated. By increasing M the efficiency compared to the traditional off-line kernel density estimator tends to one. Already for $M = 2$, it exceeds 99.1% for kernel orders and derivatives of practical interest.**

***Index Terms*— Asymptotic mean-squared error, efficiency, kernel density estimator, on-line bandwidth selection, on-line density estimator, recursive density estimator.**

I. INTRODUCTION

Let $X_1, \cdots, X_n$ denote a sequence of independent and identically distributed (i.i.d.) random variables with unknown density f. A frequently used estimator of $f(x)$ is the off-line kernel density estimator

$$\hat{f}_{OFL,n}(x) = \frac{1}{n}\sum_{i=1}^{n} h_n^{-1} K\left(\frac{x - X_i}{h_n}\right) \tag{1}$$

with K a kernel function that integrates to one, and $\{h_n\}$ a sequence of bandwidths, cf. the books by Silverman [14] or Scott [13]. A drawback of $\hat{f}_{OFL,n}(x)$ is that it requires $O(n)$ operations to update for each new observation. On the other hand, the estimator

$$\hat{f}_{REK,n}(x) = \frac{1}{n}\sum_{i=1}^{n} h_i^{-1} K\left(\frac{x - X_i}{h_i}\right) \tag{2}$$

may be computed recursively by means of the simple formula

$$\hat{f}_{REK,n}(x) = \frac{n-1}{n}\hat{f}_{REK,n-1}(x) + \frac{1}{n}h_n^{-1} K\left(\frac{x - X_n}{h_n}\right). \tag{3}$$

Recursive density estimators for the i.i.d. case were introduced by Wolverton and Wagner [17] and Yamato [18] and further treated by Wegman and Davies [16]. The properties of these estimators under different conditions of dependence are studied by, e.g., Masry [10], [11] and Tran [15].

Manuscript received January 30, 1994; revised August 26, 1994. The research of one of the authors (U. Holst) was supported by the Swedish Natural Science Research Council under Grant 9365-305.

The authors are with the Department of Mathematical Statistics, Lund Institute of Technology, S-221 00 Lund, Sweden.

IEEE Log Number 9410530.

There are a lot of connected real-time situations where rapid computation of efficient density function estimates are of great interest, e.g., recursive quantile estimation, cf. Holst [4], [5], recursive robust estimation, cf. Englund, Holst, and Ruppert [1], pattern recognition and classification, see papers by Krzyzak and Pawlak [7], [8].

Recently, Hall and Patil [3] defined on-line estimators to be those that can be updated in $O(1)$ operations for each new observation. According to (3), $\hat{f}_{REK,n}(x)$ is a member of this class. If the sequence of bandwidths are chosen optimally (for estimating $f(x)$ under squared error loss) in both (1) and (2), Hall and Patil show that the limiting relative efficiency of $\hat{f}_{REK,n}(x)$ with respect to $\hat{f}_{OFL,n}(x)$ is

$$LRE(\{\hat{f}_{REK,n}(x)\},\{\hat{f}_{OFL,n}(x)\}) = \left(\lim_{n\to\infty}\frac{E(\hat{f}_{OFL,n}(x)-f(x))^2}{E(\hat{f}_{REK,n}(x)-f(x))^2}\right)^{5/4} = 0.9295$$

for second-order kernels. (The exponent $5/4$ takes into account the convergence rate $n^{-4/5}$ for the risks, cf., e.g., Lehmann [9, ch. 5.2].)

This loss of efficiency can be regarded as the price that has to be paid for having a recursive estimator. Hall and Patil also construct other on-line estimators with limiting relative efficiency arbitrarily close to $(0.9556)^{5/4} = 0.9448$ for second-order kernels. In the present correspondence, we propose a new on-line estimator which has a higher relative efficiency.

II. Definition of the Estimators

For the development of this section, assume that a particular sequence of bandwidths $h_1,\cdots,h_n$ is given. Comparing (1) and (2) we see that $\hat{f}_{REK,n}(x)$ uses all bandwidths $h_1,\cdots,h_n$ once, whereas $\hat{f}_{OFL,n}(x)$ only uses h_n (n times). The higher efficiency of $\hat{f}_{OFL,n}(x)$ stems from the fact that it is a sum of identically distributed terms. The estimators we will construct below have the intermediate property that $h_n, h_{n-1},\cdots,h_{n-[n/M]+1}$ are used M times each. The terms of this estimator are "more similar" than those of $\hat{f}_{REK,n}(x)$.

Define a class of density estimators by

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} h_{\tau_n(i)}^{-1} K\left(\frac{x-X_i}{h_{\tau_n(i)}}\right) \tag{4}$$

where τ_n is a transformation from the set $\{1,\cdots,n\}$ into itself. This class contains as special cases $\hat{f}_{REK,n}$ in (1) ($\tau_n(i)=i$) and $\hat{f}_{OFL,n}$ in (2) ($\tau_n(i)=n$). Suppose now that τ_n is replaced by $\tau_n\circ\pi_n^{-1}$, where π_n is an arbitrary permutation of $\{1,\cdots,n\}$. The estimator in (4) is then changed to

$$\frac{1}{n}\sum_{i=1}^{n} h_{\tau_n(i)}^{-1} K\left(\frac{x-X_{\pi_n(i)}}{h_{\tau_n(i)}}\right) \tag{5}$$

and, since the sequence of observations $\{X_i\}$ is i.i.d., the distribution of $\hat{f}_n$ is unaffected. In other words, the distribution of $\hat{f}_n$ is determined by the number of times each bandwidth h_i is used. The discrete probability measure

$$p_n = \sum_{i=1}^{n} \delta_{\tau_n(i)}/n$$

contains all this information, with δ_i denoting the Dirac measure at i. We will construct an on-line estimator with

$$p_{ni} = \begin{cases} M/n, & \text{if } n-[n/M]+1 \le i \le n \\ (n-[n/M]M)/n, & \text{if } i = n-[n/M] \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

with M a positive integer and p_{ni} the relative frequency of the number of times h_i is used. Note that $M=1$ corresponds to $\hat{f}_{REK,n}$ and $M=n$ to $\hat{f}_{OFL,n}$ (even though M is considered fixed in the sequel).

A transformation that yields a distribution p_n of the form (6) is

$$\tau_n(n-i) = n-[i/M], \quad i=0,\cdots,n-1. \tag{7}$$

If τ_n is inserted into (4), the resulting estimator $\hat{f}_n$ demands $O(n)$ calculations per data value. However, there exists a permutation π_n such that $\tau_n\circ\pi_n^{-1}$ produces an on-line estimator $\hat{f}_{M,n}$. We will construct $\{\pi_n\}$ from an array I as follows: Assume $I(i)=i$ for $1\le i\le M$. If $n\ge M$, the length of I is $t_n = M(n-M+1)$. The permutation π_n will be picked from the last n elements of I. In other words, introduce a bottom indicator

$$b_n = t_n - n = (M-1)(n-M)$$

and put

$$\pi_n(i) = I(b_n+i), \quad i=1,\cdots,n. \tag{8}$$

This means that M new cells are concatenated to I each time a new observation arrives, since $t_n = t_{n-1}+M$. The concatenated elements of I are defined by putting the contents of $(I(b_{n-1}+1),\cdots,I(b_n))$ into $(I(t_{n-1}+1),\cdots,I(t_n-1))$ and letting $I(t_n)=n$. In this way, it is easy to see by induction over n that π_n defines a permutation. It remains to establish that this construction produces an on-line estimator. Notice that (7) and (8) imply $\tau_n(i)=\tau_{n-1}(i+M-1)$ and $\pi_n(i)=\pi_{n-1}(i+M-1)$ for $i=1,\cdots,n-M$. In other words, the $n-M$ pairs $\{(\tau_{n-1}(i),\pi_{n-1}(i)): i=M,\cdots,n-1\}$ are identical to $\{(\tau_n(i),\pi_n(i)): i=1,\cdots,n-M\}$. Therefore, only $M-1$ terms of $\hat{f}_{n-1}$ in (5) have to be changed when $\hat{f}_n$ is computed. The array I is actually virtual in our construction. Instead, we need to store the data, and I determines in which order to store them. Therefore, define the array $\bar{X}=(\bar{X}(1),\bar{X}(2),\cdots)$, with $\bar{X}(i)=X_{I(i)}$, which contains all the data. Combining (5), (7), and (8), we see that the resulting estimator takes the form

$$\begin{aligned}\hat{f}_{M,n}(x) &= \frac{1}{n}\sum_{i=1}^{n} h_{\tau_n(i)}^{-1} K\left(\frac{x-X_{I(b_n+i)}}{h_{\tau_n(i)}}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} h_{\tau_n(i)}^{-1} K\left(\frac{x-\bar{X}(b_n+i)}{h_{\tau_n(i)}}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} h_{[(Mn+M-n+i-1)/M]}^{-1} \\ &\quad\cdot K\left(\frac{x-\bar{X}(b_n+i)}{h_{[(Mn+M-n+i-1)/M]}}\right).\end{aligned}$$

We are now ready to formulate a recursive scheme for $\{\hat{f}_{M,n}\}$:

I Starting up ($1\le n\le M$)

1) $\bar{X}(i)=X_i, \quad i=1,\cdots,M$
2) $\hat{f}_{M,n}(x)=\hat{f}_{OFL,n}(x)$
3) $b_M=0, \quad t_M=M.$

II Induction step ($n>M$)

1) Updating $\bar{X}$:

$$\bar{X}(t_{n-1}+i) = \begin{cases} \bar{X}(b_{n-1}+i), & i=1,\cdots,M-1 \\ X_n, & i=M. \end{cases} \tag{9}$$

2) Updating $\hat{f}$:

$$\begin{aligned}\hat{f}_{M,n}(x) &= \frac{n-1}{n}\hat{f}_{M,n-1}(x) \\ &\quad - \frac{1}{n}\sum_{i=1}^{M-1} h_{[(Mn-n-i)/M]}^{-1} K\left(\frac{x-\bar{X}(b_{n-1}+i)}{h_{[(Mn-n+i)/M]}}\right) \\ &\quad + \frac{1}{n}\sum_{i=1}^{M} h_n^{-1} K\left(\frac{x-\bar{X}(t_{n-1}+i)}{h_n}\right).\end{aligned} \tag{10}$$

3) Updating b and t

$$b_n = b_{n-1} + M - 1 \quad t_n = t_{n-1} + M.$$

This means that M new cells of $\bar{X}$ are given a value in (9) for each new data value, and M terms are updated in (10). The algorithm requires storage of $\bar{X}(b_n + 1), \cdots, \bar{X}(t_n)$, which is of size $O(n)$. On the contrary, $\hat{f}_{tREK,n}(x)$ requires only $O(1)$ storage. Hence, the price for increased efficiency is enlarged memory. To our knowledge, there is no proposed on-line estimator with $O(1)$ storage and higher efficiency than $\hat{f}_{REK,n}(x)$.

III. Risk Efficiency

In this section, we will investigate the asymptotic (integrated) mean-squared error for $\hat{f}^{(s)}_{M,n}$ as an estimate of $f^{(s)}$, where s is an arbitrary nonnegative integer, and compare it to the off-line estimator $\hat{f}^{(s)}_{OFL,n}$.

For this we first need some preliminaries. Define

$$R(g) = \int g^2(u)\,du$$

for any function g and

$$\mu_j(K) = \int u^j K(u)\,du.$$

By definition, $\mu_0(K) = 1$. Define the order of the kernel as the smallest positive integer r such that $\mu_r(K) \neq 0$. We will also make the following assumptions:

i) The density f is $r + s$ times continuously differentiable at x. (If integrated mean-squared error is of concern we require f to be $r + s$ times continuously differentiable on the whole line and $R(f^{(r-s)}) < \infty$.)

ii) The kernel K satisfies

$$\int |u|^r |K(u)|\,du < \infty$$

and is s times differentiable with $R(K^{(s)}) < \infty$.

The compact support in ii) can be weakened, but then some extra conditions on f have to be imposed. We have the following result:

Theorem 1: Suppose conditions i) and ii) hold, that the kernel K is of order r, and that $h_n = cn^{-1/(2r+2s+1)}$, with c chosen optimally for both $\hat{f}^{(s)}_{M,n}$ and $\hat{f}^{(s)}_{OFL,n}$ (that is, two separate values of c). The limiting risk efficiency of $\{\hat{f}^{(s)}_{M,n}\}$ with respect to $\{\hat{f}^{(s)}_{OFL,n}\}$ is then given by

$$\begin{aligned} &LRE(\{\hat{f}^{(s)}_{M,n}\}, \{\hat{f}^{(s)}_{OFL,n}\}) \\ &\quad = \left(\lim_{n\to\infty} \frac{E(\hat{f}^{(s)}_{OFL,n}(x) - f^{(s)}(x))^2}{E(\hat{f}^{(s)}_{M,n}(x) - f^{(s)}(x))^2} \right)^{(2r+2s+1)/(2r)} \\ &\quad = \gamma_1(M)^{-(2s+1)/(2r)} \gamma_2(M)^{-1} \end{aligned} \tag{11}$$

where

$$\gamma_1(M) = \left(\frac{2r+2s+1}{r+2s+1} \right)^2 M^2 \left(1 - \left(1 - \frac{1}{M}\right)^{\frac{r+2s+1}{2r+2s+1}} \right)^2$$

and

$$\gamma_2(M) = \frac{2r+2s+1}{2r+4s+2} M \left(1 - \left(1 - \frac{1}{M}\right)^{\frac{2r+4s+2}{2r+2s+1}} \right).$$

Proof The first equality in (11) is a special case of Lehmann [9, ch. 5, Theorem 2.3], given the fact that the convergence rate is $n^{-2r/(2r+2s+1)}$ when the bandwidths are chosen as above. It remains to compare the mean-square errors to verify the second line in (11). The bias

$$b_{M,n}(c) = E(\hat{f}^{(s)}_{M,n}(x)) - f^{(s)}(x)$$

has the asymptotic expansion

$$b_{M,n}(c)^2 \sim d_1 \left(\sum_{i=1}^{n} p_{ni} h_i^r \right)^2 \tag{12}$$

with p_{ni} as in (6), $d_1 = (\mu_r(K) f^{(r+s)}(x)/r!)^2$, and $A_n \sim B_n$ meaning that $A_n/B_n \to 1$ as $n \to \infty$. Viewing the sum in (12) as a Riemann sum, we obtain

$$\begin{aligned} b_{M,n}(c)^2 &\sim d_1 c^{2r} \frac{M^2}{n^2} \left(\sum_{j=n-[n/M]}^{n} j^{-r/(2r+2s-1)} \right)^2 \\ &\sim d_1 c^{2r} \gamma_1(M) n^{-\frac{2r}{2r+2s+1}}. \end{aligned}$$

Similarly, the variance

$$v_{M,n}(c) = \mathrm{Var}(\hat{f}^{(s)}_{M,n}(x))$$

has the asymptotic expansion

$$\begin{aligned} v_{M,n}(c) &\sim d_2 n^{-1} \sum_{i=1}^{n} p_{ni} h_i^{-(2s+1)} \\ &\sim d_2 c^{-(2s+1)} \gamma_2(M) n^{-\frac{2r}{2r+2s+1}} \end{aligned}$$

with

$$d_2 = f(x) \int K^{(s)}(u)^2\,du.$$

(If integrated mean-squared error is the risk criterion, we simply modify d_1 and d_2 by integrating with respect to x over the real line.) Combining the last two displays, it follows that

$$\begin{aligned} \inf_{c>0} E\left(\hat{f}^{(s)}_{M,n}(x) - f^{(s)}(x)\right)^2 &= \inf_{c>0} \left(b_{M,n}(c)^2 + v_{M,n}(c) \right) \\ &\sim C(r,s,M) n^{-\frac{2r}{2r+2s+1}} \end{aligned} \tag{13}$$

where

$$\begin{aligned} C(r,s,M) = &\left(\frac{2r}{2s+1} \right)^{-\frac{2r}{2r+2s+1}} \frac{2r+2s+1}{2s+1} d_1^{\frac{2s+1}{2r+2s+1}} d_2^{\frac{2r}{2r+2s+1}} \\ &\cdot \gamma_1(M)^{\frac{2s+1}{2r+2s+1}} \gamma_2(M)^{\frac{2r}{2r+2s+1}}. \end{aligned}$$

The minumum is attained for

$$c \sim \left(\frac{(2s+1)\gamma_2(M) d_2}{2r \gamma_1(M) d_1} \right)^{1/(2r+2s+1)}. \tag{14}$$

In a similar way, one finds (this may be deduced from, e.g., Prakasa Rao [12, pp. 44]) that

$$\begin{aligned} \inf_{c>0} E\left(\hat{f}^{(s)}_{OFL,n}(x) - f^{(s)}(x)\right)^2 \sim &\left(\frac{2r}{2s+1} \right)^{-\frac{2r}{2r+2s+1}} \\ &\cdot \frac{2r+2s+1}{2s+1} d_1^{\frac{2s+1}{2r+2s+1}} d_2^{\frac{2r}{2r+2s+1}} n^{-\frac{2r}{2r+2s+1}}. \end{aligned} \tag{15}$$

Combining (13) and (15), we obtain (11). ■

TABLE I
LIMITING RISK EFFICIENCIES LRE$(\{\hat{f}^{(s)}_{M,n}\}, \{\hat{f}^{(s)}_{OFL\ n}\})$

		Derivative s		
		0	1	2
$r=2$	$M=1$	0.92952	0.86240	0.82990
	$M=2$	0.99765	0.99405	0.99164
	$M=3$	0.99918	0.99792	0.99707
$r=4$	$M=1$	0.95927	0.90681	0.87439
	$M=2$	0.99879	0.99661	0.99482
	$M=3$	0.99958	0.99882	0.99819
$r=6$	$M=1$	0.97135	0.92952	0.90038
	$M=2$	0.99919	0.99765	0.99628
	$M=3$	0.99972	0.99918	0.99871

Table I contains values of LRE for $M = 1, 2, 3$ and various values of r and s. Already for $M = 2$, the LRE exceeds 99.1% in all cases.

IV. BANDWIDTH SELECTION

Another important issue is bandwidth selection. The infimum in (13) is attained for a value of c that depends on f, so in reality the optimal bandwidth is unknown. Hall and Patil have shown how to estimate the optimal bandwidth from the data on-line. Their bandwidth selector could also be used for $\hat{f}_{M,n}$, since the expression for the optimal bandwidth is essentially the same.

More precisely, recall that the minimum in (13) is attained for a c given by (14). The optimal bandwidth is easiest to estimate when the risk criterion is integrated mean-squared error. Then

$$d_1 = \mu_r(K)^2 \int f^{(r+s)}(x)^2\,dx/(r!)^2$$

and

$$d_2 = \int K^{(s)}(u)^2\,du.$$

The only unknown quantity in (14) is

$$R(f^{(r+s)}) = \int f^{(r+s)}(x)^2\,dx$$

regardless of M. Hall and Patil construct an on-line estimator of $R(f^{(2)})$. A generalization of their estimator to arbitrary r and s takes the form

$$\hat{R}_n = (-1)^{r+s} m^{-1} \left(n - \frac{1}{2}(m+1)\right)^{-1} \cdot \sum_{i=1}^{m} \sum_{j=i+1}^{n} l_j^{-2r-2s-1} K_1^{(2r+2s)}\left(\frac{X_j - X_{j-i}}{l_j}\right)$$

provided $n \geq m+1$, where m is a fixed positive integer, K_1 is a kernel of order r_1, and $\{l_j\}$ is a new sequence of bandwidths. (For $n \leq m$ we may, for instance, take $\hat{R}_n$ an arbitrary constant.) The estimator $\hat{R}_n$ is constructed so that it can be computed from $\hat{R}_{n-1}$ in $O(1)$ steps. To justify that $\hat{R}_n$ is a reasonable estimator of $R(f^{(r+s)})$, notice that

$$R(f^{(r+s)}) = (-1)^{r-s} E(f^{(2r+2s)}(X))$$

which follows from integration by parts. Then notice that

$$\hat{R}_n = \frac{(-1)^{r+s}}{n-1} \sum_{j=2}^{n} \hat{f}^{(2r+2s)}_{nj}(X_j)$$

where

$$\hat{f}_{nj}(x) = \frac{n-1}{n-(m+1)/2} \cdot \frac{1}{m} \sum_{i=1}^{(j-1)\wedge m} l_j^{-1} K_1\left(\frac{x - X_{j-i}}{l_j}\right).$$

When $j > m$, this expression is asymptotically equivalent to

$$\frac{1}{m} \sum_{i=1}^{m} l_j^{-1} K_1\left(\frac{x - X_{j-i}}{l_j}\right)$$

as n tends to infinity. Hall and Marron [2] and Jones and Sheather [6] consider the problem of estimating integrated squared density derivatives in more detail.

Suppose now that

i') the first $r+s+r_1$ derivatives of f are bounded, continuous, and integrable.

iii)The kernel K_1 is $2r+2s$ times differentiable, with

$$\int |u|^{r_1} |K_1^{(r+s)}(u)|\,du < \infty$$

and $R(K_1^{(2r+2s)}) < \infty$.

Then

$$E\hat{R}_n - R(f^{(r+s)}) \sim \mu_{r_1}(K_1) \int f^{(r+s)}(x) f^{(r+s-r_1)}(x)\,dx \cdot \sum l_j^{r_1}/(n(r_1)!)$$

and

$$\mathrm{Var}(\hat{R}_n) \sim R(K_1^{(2r+2s)}) R(f) \sum l_j^{-(4r+4s+1)}/(mn^2).$$

Notice that the leading bias term of $\hat{R}_n$ vanishes when r_1 is odd. Choosing $l_j = c_1 j^{-1/(2r_1+4r+4s+1)}$ for some fixed constant $c_1 > 0$ gives the optimal rate of convergence (when r_1 is even)

$$\hat{R}_n - R(f^{(r+s)}) = O_p\left(n^{-r_1/(2r_1+4r+4s+1)}\right).$$

Now define $\hat{c}_j$ by (14), except that $R(f^{(r+s)})$ is replaced by $\hat{R}_{j-1}$ in the definition of d_1, and let

$$\hat{h}_{n,i} = \hat{c}_{\pi_n(i)}[(Mn + M - n + i - 1)/M]^{-1/(2r+2s+1)}$$

with $\pi_n(i)$ defined in (8). We may now define a "plug-in" version of $\hat{f}_{M,n}$ as

$$\hat{f}_{M,n}(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{h}_{n,i}^{-1} K\left(\frac{x - X_{\pi_n(i)}}{\hat{h}_{n,i}}\right). \tag{16}$$

Observe that when $h_n = cn^{-1/(2r+2s+1)}$, $\hat{f}_{M,n}(x)$ is defined by replacing c with $\hat{c}_{\pi_n(i)}$ in the ith term of the definition of $\hat{f}_{M,n}$. If follows from the above considerations that $\hat{f}_{M,n}$ is an on-line estimator, provided we store the numbers $\hat{c}_j, j = 1, \cdots, n$.

Theorem 2: Given i'). ii), and iii), the on-line, plug-in kernel density estimator $\hat{f}_{M,n}(x)$ defined in (16) is asymptotically normal in the sense that

$$(\hat{f}_{M,n}(x) - b_{M,n}(c))/\sqrt{v_{M,n}(c)}$$

converges in distribution to the standard normal distribution as $n \to \infty$, where $b_{M,n}(c)$ and $v_{M,n}(c)$ are the bias and variance for $\hat{f}_{M,n}(x)$ when $h_n = cn^{-1/(2r+2s-1)}$ and c is defined in (14), with

$$d_1 = \mu_r(K)^2 R(f^{(r+s)})/(r!)^2$$

and $d_2 = R(K^{(s)})$.

Proof: Observe that $\tilde{f}_{M,n}(x)$ may be written as

$$\tilde{f}_{M,n}(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{h}_{n,\pi_n^{-1}(i)}^{-1} K\left(\frac{x - X_i}{\hat{h}_{n,\pi_n^{-1}(i)}}\right). \tag{17}$$

By definition

$$\hat{h}_{n,\pi_n^{-1}(i)} = \hat{c}_i\left([(Mn + M - n + \pi_n^{-1}(i) - 1)/M]\right)^{-1/(2r+2s+1)}$$

so it depends only on $X_1, \cdots, X_{i-1}$. Let

$$U_{n,i} = K\left((x - X_i)/\hat{h}_{n,\pi_n^{-1}(i)}\right)/\hat{h}_{n,\pi_n^{-1}(i)}$$

be the ith term in (17) and denote $u_{n,i} = E(U_{n,i}|X_1, \cdots, X_{i-1})$ and $V_{n,i} = U_{n,i} - u_{n,i}$. Note that $\{V_{n,i}\}$ are martingale differences. As in [3, Appendix (ii)]) one shows that

$$\frac{1}{\sqrt{v_{M,n}(c)\, n}} \sum_{i=1}^{n} (u_{n,i} - b_{M,n}(c)) \xrightarrow{p} 0.$$

Further, using a Martingale central limit theorem, it follows that

$$\frac{1}{\sqrt{v_{M,n}(c)\, n}} \sum_{i=1}^{n} V_{n,i} \xrightarrow{d} N(0,1). \qquad \blacksquare$$

REFERENCES

[1] J.-E. Englund, U. Holst, and D. Ruppert, "Recursive M-estimation of location and scale for dependent sequences," *Scand. J. Statist.*, vol. 15, pp. 147–159, 1988.

[2] P. Hall and S. Marron, "Estimation of integrated squared density derivatives," *Stat. Prob. Lett.*, vol. 6, pp. 109–115, 1987.

[3] P. Hall and P. Patil, "On the efficiency of on-line density estimators," *IEEE Trans. Inform. Theory*, vol. 40, no. 5, pp. 1504–1512, Sept. 1994.

[4] U. Holst, "Recursive estimation of quantiles using recursive kernel density estimators," *Sequential Anal.*, vol. 6, pp. 219–237, 1987.

[5] ——, "Recursive M-estimators of location," *Commun. Statist.–Theory and Methods*, vol. 16, pp. 2201–2226, 1987.

[6] C. Jones and S. Sheather, "Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives," *Stat. Prob. Lett.*, vol. 11, pp. 511–514, 1991.

[7] A. Krzyzak and M. Pawlak, "Universal consistency results for Wolverton–Wagner regression function estimate with application in discrimination," *Probl. Contr.–Inform. Theory*, vol. 12, pp. 32–42, 1983.

[8] ——, "Almost everywhere convergence of a recursive regression function estimate and classification," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 91–94, 1984.

[9] E. Lehmann, *Theory of Point Estimation*. Wadsworth and Brooks, 1991.

[10] E. Masry, "Recursive probability density estimation for weakly dependent stationary processes," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 254–267, 1986.

[11] ——, "Almost sure convergence of recursive density estimators for stationary mixing processes," *Statist. Probab. Lett.* vol. 5, pp. 249–254, 1987.

[12] B. P. Rao, *Nonparametric Functional Estimation*. New York: Academic Press. 1983.

[13] D. Scott, *Multivariate Density Estimation. Theory, Practice, and Visualization*. New York: Wiley, 1992.

[14] B. Silverman, *Density Estimation for Statistics and Data Analysis*. London, UK: Chapman and Hall, 1986.

[15] L. Tran, "Recursive density estimation under dependence," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1103–1108, 1989.

[16] E. Wegman and H. Davies, "Remarks on some recursive estimators of a probability density," *Ann. Statist.*, vol. 7, pp. 316–327, 1979.

[17] C. Wolverton and T. Wagner, "Asymptotically optimal discriminant functions for pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 258–265, 1969.

[18] H. Yamato, "Sequential estimation of a continuous probability density function and mode," *Bull. Math. Statist.*, vol. 14, pp. 1–12, 1971.

On the Relation Between Filter Maps and Correction Factors in Likelihood Ratios

Ravi R. Mazumdar, *Senior Member, IEEE*,
and Arunabha Bagchi, *Senior Member, IEEE*

***Abstract*— The robust form of the likelihood ratio for a signal in the presence of white noise has an additional term in the exponent called the correction factor which corresponds to the trace of the conditional covariance of the signal given the observations. In this correspondence we show that this correction term is nothing but the trace of the symmetrized Fréchet derivative of the nonlinear filter map and hence the likelihood ratio can be completely represented in terms of the observations and the filter map.**

***Index Terms*— Likelihood ratios, nonlinear filters, correction factors, white noise, Fréchet derivatives.**

I. INTRODUCTION

In this correspondence we show the relationship between the so-called "correction factors" associated with the robust form of the likelihood ratio for random signals in white noise and the nonlinear filter map associated with the filtering problem. In particular, we show that the correction term is the trace of the symmetrized Fréchet derivative of the nonlinear filter map viewed as a map on the observations as a mapping from $(L_2(0,T);\Re^n)$ into itself.

It is convenient to begin with a discussion of the issue in the context of likelihood ratios (or Radon–Nikodym derivatives) in the Gaussian context and for convenience we assume the processes take values in $\Re$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space which carries a filtration $\{\mathcal{F}_t\} \subset \mathcal{F}$. Let $\{S_t\}_{t\in[0,T]}$ be a zero-mean Gaussian process adapted to $\{\mathcal{F}_t\}$ which satisfies

$$\mathbb{E}[\int_0^T |S_t|^2 dt] < \infty.$$

Let $\{W_t\}$ be a $\mathcal{F}_t$ Brownian motion independent of $\mathcal{F}_t^S$ where $\mathcal{F}_t^S$ denotes the filtration generated by $\{S_t\}$. Define the so-called "observation" process $\{Y_t\}$ given by

$$dY_t = S_t dt + dW_t; t \in [0, T] \tag{1}$$

Then the classical formula for the Radon–Nikodym derivative (RND) of the measure induced by Y denoted by μ_Y with respect to (w.r.t.) μ_W the standard Wiener measure is given by (see [6])

$$\frac{d\mu_Y}{d\mu_W}(Y) = \exp\{\oint_0^T \hat{S}_u dY_u - \frac{1}{2}\int_0^T |\hat{S}_u|^2 du\} \tag{2}$$

where $\hat{S}_t = \mathbb{E}[S_t/\mathcal{F}_t^Y]$, $\mathcal{F}_t^Y$ is the filtration generated by the observations. $\hat{S}_t$ is called the filtered process and the integral $\oint$ in (2) denotes the Ito integral.

It is important to point out that the result of Kailath [6] holds not just for Gaussian signals $\{S_t\}$ but also for any signal satisfying

$$\mathbb{E}[\int_0^T |S_t|^2 dt] < \infty.$$

Manuscript received November 16, 1993; revised November 10, 1994.

R. R. Mazumdar is with INRS-Télécommunications, Université du Québec, Ile des Soeurs, P.Q. H3E 1H6, Canada.

A. Bagchi is with the Department of Applied Mathematics, University of Twente, 7500AE Enschede, The Netherlands.

IEEE Log Number 9410531.